DOCUMENT RESUME

ED 073 125 TM 002 368

AUTHOR Meredith, Keith E.; Sabers, Darrell L.
TITLE Using Item Data for Evaluating Criterion Reference Measures with an Empirical Investigation of Index Consistency.
PUB DATE 16 Nov 72
NOTE 12p.; Paper presented at Rocky Mountain Educational Research Association, November 16, 1972

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS *Criterion Referenced Tests; Factor Analysis; *Item Analysis; Research Methodology; Speeches; Tables (Data); *Test Reliability; *Test Reviews; *Test Validity

AESTRACT

Data required for evaluating a Criterion Referenced Measurement (CRM) is described with a matrix. The information within the matrix consists of the "pass-fail" decisions of two CRMs. By differentially defining these two CRMs, different concepts of reliability and validity can be examined. Indices suggested for analyzing the matrix are listed with their formula. Using the National Norm Group for the Primary Test of Economic Understanding, three questions were investigated: (1) What indices are consistent across samples?; (2) What indices are giving independent information?; and (3) Are these same indices consistent when sample sizes vary? The procedure and analysis used to obtain answers to the questions are given, and five tables provide the study data. (DB)

ED 073125

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY

USING ITEM DATA FOR EVALUATING CRITERION REFERENCE MEASURES WITH AN EMPIRICAL INVESTIGATION OF INDEX CONSISTENCY

Keith E. Meredith and Darrell L. Sabers

University of Arizona

A paper read at Rocky Mountain Educational Research Association

November 16, 1972

TM 002 368

With the current trends towards accountability, management by objectives, and objective based evaluation measurement of specific behavioral objectives will undoubtedly increase. There are two approaches to reporting an individual's performance on an objective. Norm referenced measurement (NRM) provides for assessment of objectives; however, its theory is based upon comparison of an individual's performance to the performance of some defined population. Criterion referenced measurement (CRM) also provides for assessment of objectives; however, its theory is based upon comparison of an individual's performance to an established criterion. Information concerning this individual's rank in relation to a population is basically of no interest.

The value of each theoretical base has been established in other discussions. Procedures for investigating and improving NRM are greatly developed partially as a result of longevity. CRM lacks such investigative procedures due to its more recent arrival.

Psychometricians have attempted to apply investigative techniques developed for NRM to the newer CRM. Obviously, in view of their theoretical differences, many of these techniques have been found wanting. New procedures have been developed for specific cases of CRM: however, their generalizability has not been investigated. What is overdue is a thorough

look at what concerns are appropriate for CRM and what techniques can best be used to investigate these concerns within CRM.

Validity and reliability, in that order, are the two "qualities of any test" that receive the majority of attention in any measurement discussion. Although one seldom establishes the validity of a NR test in the absence of reliability data, evidence is usually desired to establish both properties of an achievement test. The same types of evidence should be required of CR measures.

Valid arguments have been made to suggest the CRM and NRM differ on one important aspect--that of a total score. While a total score is usually desired in the case of a NR test, it will seldom be meaningful in the case of CRM. In fact, CRM produces one score for each objective measured, and a "total" score simply represents a summary of the performance on a number of tasks on one objective. When only one or two items are available per objective, statistical evaluation of CRM becomes similar to item analysis in NRM. Thus, a treatment of validity and reliability for CRM deals with item analysis.

If we are to use item analysis procedures for evaluating a CRM then we must determine what type of score is produced by the item. Scores may be dichotomous or multi-score. Since the decision which is usually made as a result of a CRM is whether an individual has mastered an objective, the multi-score item would eventually result in a dichotomous item for purposes of decision-making.

A CRM ("item") can make two types of incorrect decisions in this "pass-fail" situation. Hambleton and Novick (1972) referred to these errors as false positives and false negatives. A false positive error occurs when the CRM identifies the individual as having "passed" or mastered the objective

when, in fact, he has not. In the situation where the CRM identifies the individual as having "failed" the object when, in fact, he has achieved the criterion then the CRM has made a false negative error.

Reliability, then, can be thought of as the CRM's ability to consistently make the same decision. Validity is the CRM's ability to make the appropriate decision. "Appropriate" means that neither a false positive or a false negative decision error has been committed in reference to the objective being measured.

As in NRM, to examine these aspects we must look at results from groups of examinees. The adequacy of a CRM will be determined by its ability to discriminate consistently and appropriately over a large number of cases.

Reliability and validity procedures in NRM are based upon the assumption that variability exists in the construct being measured. If this assumption is not met both reliability and validity indices are correspondingly low.

In CRM this assumption is not a viable one to make. In many instructional situations this assumption will not even be approximated. In fact, this heterogeneity of attainment is contrary to the philosophy which gave rise to CRM.

This difference between NRM and CRM is the basic criticism of the work of Livingston (1972). Livingston devised a reliability coefficient based on the degree of deviation from a criterion score (rather than deviations from the mean as in NRM). But CRM is concerned only with the accuracy of the "pass-fail" decision and is relatively unconcerned with a person's attainment above or below this cut-off point. Eventually this may be of interest to give an indication of how much additional training the examinee needs if he has not yet met the established criterion.

Carver (1970) proposed two procedures to assess reliability of a CRM. For a single form he suggests comparing the percentage identified as meeting the objective in one group to the percentage identified as meeting the objective in another similar group. For parallel-form reliability he recommends using one group and comparing the percentages identified as having criterion on the two parallel forms.

Similar logic in NRM would require the computation of a reliability coefficient using two different groups of individuals. We must know how two CRMs, identical or parallel, identify the same examinee in regard to his attainment of the objective.

Various authors (Popham and Husek, 1969; Helmstadter, 1972; Cox and Vargis, 1966) have used item analysis procedure for determining content validity. A CRM is administered previous to a unit of instruction and again following the unit. The assumption is that if the item discriminates between the pre- and posttests then it is reliable and valid for the content of that course.

This assumption is acceptable when the item does discriminate. However, when the item does not discriminate we do not know if the item is unreliable or an invalid measure of the objective; whether the objective is inappropriate for the course of instruction, or whether the instruction did not teach the objective. This pre-post procedure should be used to examine content validity after the items have been shown to be both reliable and valid for the objective being measured.

The data required for evaluating a CRM can be described with a matrix. The information within the matrix consists of the "pass-fail" decisions of two CRMs. By differentially defining these two CRMs we can examine different concepts of reliability and validity. This matrix is shown below.

		CRM #1	
		Pass	Fail
CRM #2	Pass	A	B
	Fail	C	D

By defining the two CRM's as being the same measure we can examine test-retest reliability. If the two CRM's are different measures of the same objective, parallel forms reliability can be expressed. Validity of one CRM can be determined if the other CRM is a criterion measure.

Various indices have been suggested for analyzing this matrix. These indices are listed, with their formula, on the following page.

Given all of these indices two questions arise. What information is being derived from these indices, and which indices are giving this information most accurately and consistently?

We would expect that, given a specific pool of pairs of items, each pair a measure of a common objective, an index should rank these measures in the same order across samples from a population. To the degree that an index can accomplish this, it can be considered a consistent measure.

By determining whether any of these indices tend to group together, i.e. provide similar information, we can determine what information is being gained.

To exhibit accuracy an index should provide one type of information and be independent of other groups of indices. If variables are confounded then one is not certain what information is derived.

Index	Formula
Phi	$\frac{AD - BC}{\sqrt{(A+C)(B+D)(A+B)(C+D)}}$
Lamda	$\frac{1 + [\Sigma p\alpha\alpha - \frac{1}{2}(PM\cdot + P\cdot M)/1 - \frac{1}{2}(PM\cdot + P\cdot M)]}{2}$
Lamda K	$\frac{[(B+C+D)/3]+D}{[(B+C+D)/3] + B + C + D}$
Difficulty of CRM #1	$\frac{A + C}{N}$
Difficulty of CRM #2	$\frac{A + B}{N}$
Average Difficulty	$\frac{\text{Difficulty of CRM\#1} + \text{Difficulty of CRM\#2}}{2}$
Objective Difficulty	$\frac{A}{N}$
Deviation between Expected Objective Difficulty and Obtained Objective Difficulty	$[(A+C)(A+B)] - \frac{A}{N}$
Agreement between CRMs	$\frac{A + D}{N}$
Difference between Error Cells	$\frac{C - B}{N}$
Objective Difficulty + Lamda	$\frac{A}{N}$ + Lamda
Objective Difficulty + Lamda K	$\frac{A}{N}$ + Lamda K

Methodology

Data Base

National Norm Group for the Primary Test of Economic Understanding (Davison and Kilgore, 1971).

This test consists of 64 true-false items over 32 concepts. Each concept has two items measuring it with one item keyed true and one item keyed false. Each item can therefore be considered a CRM over the underlying concept. This pair of CRMs can then be placed in the matrix format indicated earlier. The national norm group consisted of 166 classrooms from 20 states.

Procedure and Analysis

Three principal questions must be investigated: 1. What indices are consistent across samples; 2. what indices are giving independent information; and are these same indices consistent when sample sizes vary?

To answer the first two questions, classrooms were assigned to subgroups. The twelve indices listed on page 6 were computed for each of the 32 concepts on PTEU for each subgroup. Each index for a concept would, therefore, be replicated in each subgroup. In order to demonstrate consistency an index should rank these 32 concepts the same for all subgroups. If we call the concepts our sample and the two values of the index derived from two different subgroups our two variables X and Y, then the correlation between X and Y should be higher for every combination of subgroups.

All of the indices computed were entered into a factor analysis. If all of the combinations of subgroups result in similar correlations then all replications of the same index should load consistently on the same factor. By taking an average of the loadings of the replications of an index on a factor and also computing a standard deviation one can examine the strength

of the loading on that factor (mean of the replications) and the consistency across subgroups (standard deviation of the replications). The aspect of independent information can be investigated by comparing the mean loadings across factors. An index which provides independent information would have a high mean loading on one factor and low mean loadings on the remaining factors. Consistency of an index would be evidenced by small standard deviations of the loadings across subgroup.

Consistency of indices across varying sample sizes (question 3) was determined by performing the analysis discussed above on four major groups. The classrooms were randomly assigned to a particular subgroup within a major group. The subgroups within a major group were all of equal size, but size differed across the major groups.

The first three major groups consisted of five subgroups each. The sizes of these subgroups were one, five, and ten classrooms respectively. The fourth major group was formed with two subgroups, each made up of thirty classrooms.

In terms of "students" the average size of subgroups within each major group was 30, 150, 300 and 800 "students" respectively.

Table 1

Group Assignment

Major Group I	Major Group II	Major Group III	Major Group IV
5 subgroups	5 subgroups	5 subgroups	2 subgroups
Each subgroup consisted of one classroom	Each subgroup consisted of five classrooms	Each subgroup consisted of ten classrooms	Each subgroup consisted of thirty classrooms

Consistency of indices across sample sizes would be demonstrated by similar factor structurs across the major groups.

Table 2

Means and Standard Deviations of Factor Loadings
For Each Index for Group I

(Each mean is based on five subgroups;
Each subgroup consists of one classroom)

Index	Factor 1		Factor 2		Factor 3	
	Mean	S.D.	Mean	S.D.	Mean	S.D.
Phi	.15	.17	.06	.25	.35	.31
Lamda	.06	.09	.03	.15	.53	.29
Lamda K	-.19	.22	.04	.10	.54	.26
Difficulty of CRM#1	-.14	.55	.43	.10	.05	.20
Difficulty of CRM#2	.07	.14	-.57	.16	.31	.21
Average Difficulty	.30	.28	.41	.14	-.19	.16
Objective Difficulty	.32	.18	-.58	.18	.16	.24
Deviation between Expected Objective Difficulty and Obtained Objective Difficulty	.16	.19	-.17	.29	-.23	.29
Agreement between CRMs	.64	.08	-.06	.13	.00	.17
Difference between Error Cells	-.08	.27	.61	.18	-.33	.13
Objective Difficulty + Lamda	.72	.08	-.07	.13	.09	.24
Objective Difficulty + Lamda K	.73	.07	-.02	.13	.02	.21

Table 3

Means and Standard Deviations of Factor Loadings
For Each Index for Group II

(Each mean is based on five subgroups;
Each subgroup consists of five classrooms)

Index	Factor 1		Factor 2		Factor 3	
	Mean	S.D.	Mean	S.D.	Mean	S.D.
Phi	.34	.08	-.05	.06	.28	.61
Lamda	.08	.16	.06	.21	.64	.33
Lamda K	-.25	.32	.04	.18	.61	.32
Difficulty of CRM#1	-.49	.10	-.63	.15	.28	.06
Difficulty of CRM#2	-.20	.18	.78	.13	.30	.05
Average Difficulty	.20	.02	.92	.01	-.01	.04
Objective Difficulty	.36	.04	.88	.02	.06	.05
Deviation between Expected Objective Difficulty and Obtained Objective Difficulty	-.22	.09	.00	.05	-.65	.09
Agreement between CRMs	.93	.02	.07	.06	.08	.04
Difference between Error Cells	-.07	.07	.94	.01	-.04	.04
Objective Difficulty + Lamda	.91	.02	.07	.05	.10	.04
Objective Difficulty + Lamda K	.93	.02	.07	.06	.08	.04

Table 4

Means and Standard Deviations of Factor Loadings
For Each Index for Group III

(Each mean is based on five subgroups;
Each subgroup consists of ten classrooms)

Index	Factor 1		Factor 2		Factor 3	
	Mean	S.D.	Mean	S.D.	Mean	S.D.
Phi	.37	.24	-.02	.14	.50	.29
Lamda	.03	.09	.09	.16	.71	.36
Lamda K	-.28	.27	.05	.14	.68	.33
Difficulty of CRM#1	-.46	.05	-.79	.08	.17	.06
Difficulty of CRM#2	-.07	.12	.91	.02	.27	.05
Average Difficulty	.20	.04	-.94	.02	-.04	.08
Objective Difficulty	.40	.04	.89	.02	.09	.04
Deviation between Expected Objective Difficulty and Obtained Objective Difficulty	-.34	.09	-.03	.13	-.71	.11
Agreement between CRMs	.95	.01	.06	.04	.09	.03
Difference between Error Cells	-.12	.05	.96	.01	-.08	.05
Objective Difficulty + Lamda	.94	.01	.07	.05	.06	.10
Objective Difficulty + Lamda K	.95	.01	.06	.04	.05	.09

Table 5

Means and Standard Deviations of Factor Loadings
For Each Index for Group IV

(Each mean is based on two subgroups;
Each subgroup consists of thirty classrooms)

Index	Factor 1		Factor 2		Factor 3	
	Mean	S.D.	Mean	S.D.	Mean	S.D.
Phi	.90	.01	.13	.07	.03	..04
Lamda	.38	.02	-.89	.01	-.14	.02
Lamda K	.39	.01	-.89	.01	-.11	.02
Difficulty of CRM#1	-.01	.00	.88	.01	.47	.01
Difficulty of CRM#2	.16	.02	.74	.00	-.63	.00
Average Difficulty	.09	.01	.99	.00	-.07	.01
Objective Difficulty	.28	.00	.95	.00	-.01	.15
Deviation between Expected Objective Difficutly and Obtained Objective Difficulty	.84	.02	.35	.04	.09	.04
Agreement between CRMs	.68	.03	.61	.03	-.32	.01
Difference between Error Cells	-.14	.02	.19	.01	.96	.00
Objective Difficulty + Lamda	.83	.02	.30	.06	-.37	.02
Objective Difficulty + Lamda K	.64	.02	.70	.01	-.28	.01